DATASETS:

1. US Health Insurance Dataset from Kaggle
   https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset

2. Dermatology Dataset (Multi-class classification) from Kaggle
   https://www.kaggle.com/datasets/olcaybolat1/dermatology-dataset-classification/data

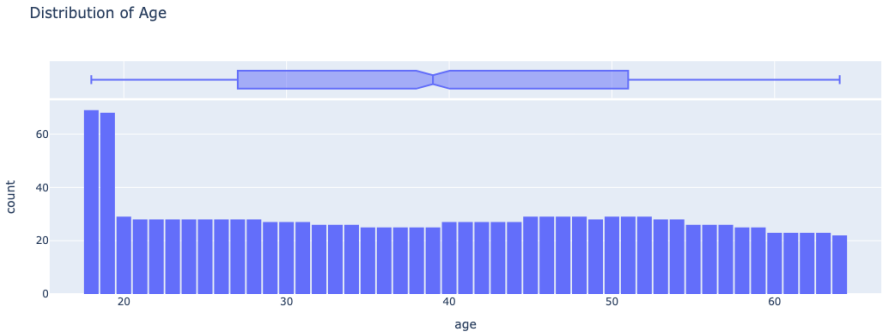
Market segment analysis (Datasets 1)

The dataset seems to be focused on factors that could influence health insurance costs for individuals. It includes both demographic (age, sex, region) and health-related factors (BMI, smoking status), as well as information on dependents, which are all relevant in determining insurance premiums or analyzing health risks.

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

1338 rows × 7 columns

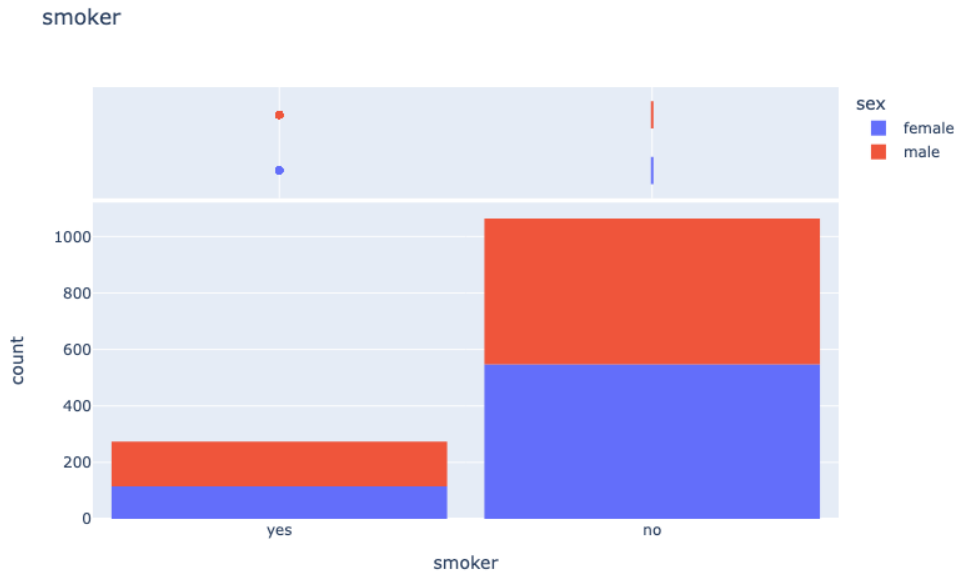
combined box plot and histogram of the distribution of age within a dataset



1. **Histogram Analysis**:
   - The histogram shows a unimodal distribution with the highest frequency of individuals in the youngest age bracket.
   - There's a high concentration of individuals in their 20s, and the frequency gradually decreases as age increases.

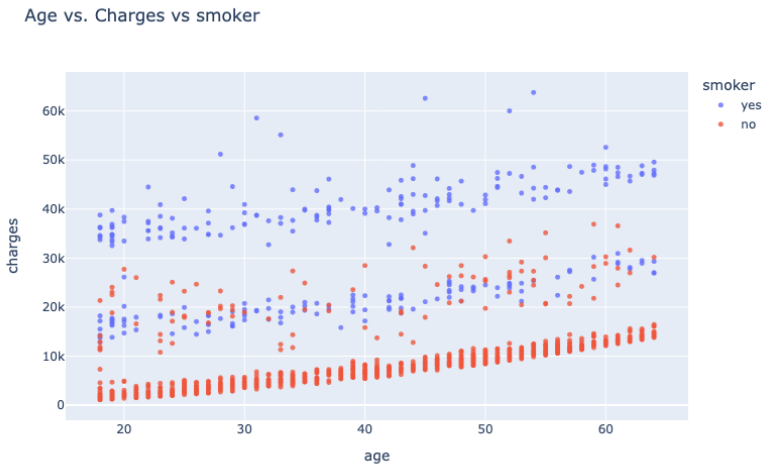
- There are fewer individuals in the dataset who are older, with a very low frequency of individuals over 60 years of age.

Bar chart with a count plot



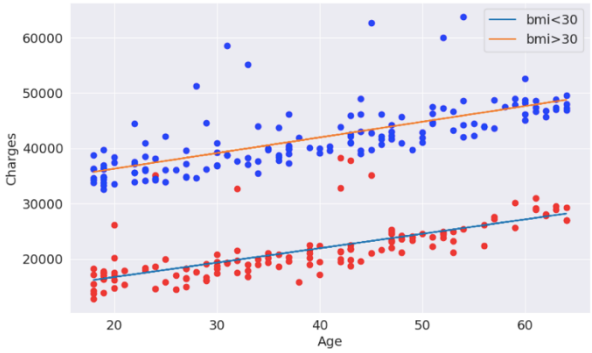
1. There are significantly more non-smokers than smokers in this dataset.
2. For both smokers and non-smokers, the distribution between males and females is quite similar, with a slightly higher count of males in both categories.

Scatter plot comparing age, charges, and smoking status



1. **Age and Charges Correlation**: The plot suggests a positive correlation between age and insurance charges; as age increases, the charges tend to increase as well. This trend is common in insurance pricing, where older individuals are often charged more due to higher associated health risks.

2. **Impact of Smoking**: There are two distinct clusters of data points, which likely represent smokers and non-smokers. The cluster with higher charges at any given age represents smokers ('yes'), which indicates that smokers are charged significantly more for insurance than non-smokers ('no'). This is a typical finding, as smoking is associated with higher health risks and therefore higher insurance costs.



3. **Trend Lines**: There are two trend lines that appear to fit the data for each category.

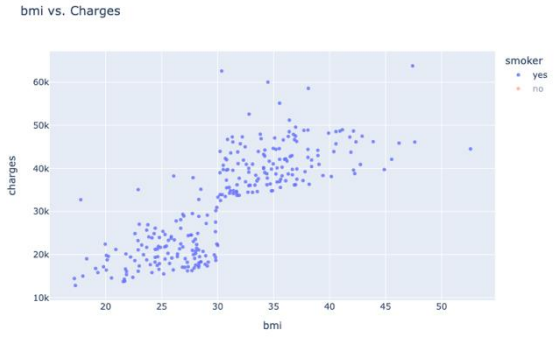
Compare BMI to insurance charges

For Individuals with BMI ≤ 30:

- Among non-smokers (red points in the first plot), those with a BMI ≤ 30 seem to have a fairly uniform distribution of charges that don't show a strong dependency on BMI within this range.
- Smokers (blue points in the first plot) with a BMI ≤ 30 also display an increase in charges, but the range is more variable and generally higher than non-smokers.

For Individuals with BMI > 30:

- Non-smokers with a BMI > 30 (red points) do not show a significant increase in charges compared to those with a BMI ≤ 30, suggesting that while BMI is a factor, its impact on charges is not as pronounced for non-smokers.
- Smokers with a BMI > 30 (blue points) show a trend of even higher charges, which indicates that a higher BMI exacerbates the

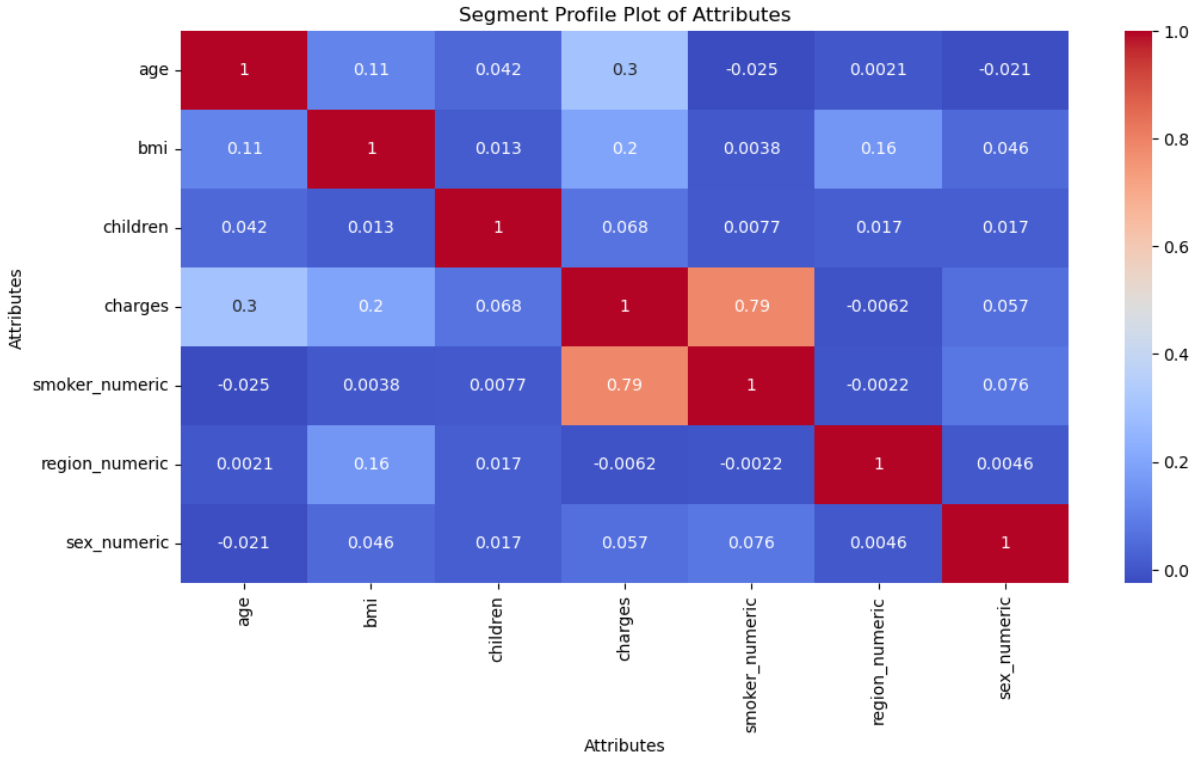


cost of insurance for smokers. The combination of being a smoker and having a high BMI seems to result in the highest insurance charges.

Overall Trends:

- The plots collectively suggest that smoking status has a more substantial impact on insurance charges than BMI alone.
- Individuals with a higher BMI (>30), particularly smokers, are at the higher end of insurance charges, underscoring the combined effect of smoking status and higher BMI as significant factors in determining insurance costs.
- The clustering of charges for non-smokers remains tighter and lower across the range of BMI compared to smokers, indicating a less steep correlation between BMI and charges for non-smokers.

Heatmap of a correlation matrix



Strong Positive Correlation: The most visually prominent element is the strong positive correlation between **smoker_numeric** and **charges** (0.79), indicating that smoking status is highly associated with higher insurance charges

OLS regression results (data 1 including smoker)

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.755
Model:	OLS	Adj. R-squared:	0.750
Method:	Least Squares	F-statistic:	164.8
Date:	Fri, 01 Mar 2024	Prob (F-statistic):	1.36e-79
Time:	16:52:49	Log-Likelihood:	-2758.7
No. Observations:	274	AIC:	5529.
Df Residuals:	268	BIC:	5551.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.117e+04	985.505	-11.331	0.000	-1.31e+04	-9226.329
age	263.4997	25.253	10.434	0.000	213.780	313.219
bmi	1453.6802	57.276	25.380	0.000	1340.913	1566.448
children	216.8542	303.946	0.713	0.476	-381.571	815.279
smoker_numeric	-1.117e+04	985.505	-11.331	0.000	-1.31e+04	-9226.329
region_numeric	-295.0558	331.593	-0.890	0.374	-947.913	357.802
sex_numeric	-346.2074	717.154	-0.483	0.630	-1758.179	1065.765

Omnibus:	60.164	Durbin-Watson:	1.901
Prob(Omnibus):	0.000	Jarque-Bera (JB):	143.018
Skew:	1.037	Prob(JB):	8.79e-32

1. **Model Fit**: An R-squared value of 0.755 indicates a strong model fit, explaining approximately 75.5% of the variance in insurance charges.
2. **Significant Predictors**: Age and BMI have a statistically significant positive relationship with insurance charges.
3. **Children**: The number of children has a positive coefficient but is not statistically significant (p-value = 0.476).
4. **Region and Sex**: Both region_numeric and sex_numeric show no significant impact on insurance charges (p-values = 0.374 and 0.630, respectively).
5. **Model Diagnostics**:
    - The model is statistically significant as a whole, indicated by the F-statistic.
    - The Durbin-Watson statistic of 1.901 suggests no significant autocorrelation concerns.
    - Tests for normality indicate significant skewness and kurtosis in the residuals, questioning the p-value reliability.
6. **Coefficient Implications**:
    - For age, a one-unit increase corresponds to an approximate increase in charges by 263.50.
    - The BMI coefficient suggests a one-unit increase in BMI corresponds to an increase in charges by 1453.68.

OLS regression results (data 2 not including smoker)

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.417
Model:	OLS	Adj. R-squared:	0.415
Method:	Least Squares	F-statistic:	151.5
Date:	Fri, 01 Mar 2024	Prob (F-statistic):	2.09e-121
Time:	16:52:49	Log-Likelihood:	-10477.
No. Observations:	1064	AIC:	2.097e+04
Df Residuals:	1058	BIC:	2.100e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2236.4125	812.388	-2.753	0.006	-3830.486	-642.339
age	264.5855	10.072	26.269	0.000	244.822	284.349
bmi	18.4778	23.706	0.779	0.436	-28.039	64.995
children	587.2669	115.565	5.082	0.000	360.503	814.030
smoker_numeric	2.153e-13	8.64e-14	2.490	0.013	4.57e-14	3.85e-13
region_numeric	-461.9849	127.882	-3.613	0.000	-712.916	-211.053
sex_numeric	-526.7463	281.474	-1.871	0.062	-1079.057	25.564

Omnibus:	710.730	Durbin-Watson:	2.053
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5704.085
Skew:	3.191	Prob(JB):	0.00

...

1. **Model Fit**: R-squared is 0.417, suggesting a moderate fit with the model explaining about 41.7% of variance in insurance charges.
2. **Significant Predictors**:
    - Age is a significant predictor with a positive coefficient, indicating charges increase with age.
    - Children are also a significant predictor, with more children associated with higher charges.
3. **Non-Significant Predictors**:
    - BMI has a positive coefficient but is not statistically significant (p-value = 0.436).
    - The coefficient for smoker status is negligible and not consistent with typical interpretations of its impact on charges.
4. **Other Observations**:
    - Region has a significant negative coefficient, indicating some regions are associated with lower charges.
    - Sex has a negative coefficient, suggesting a possible but weak association with lower charges for males.
5. **Coefficient Interpretation**:
    - Coefficients indicate the change in charges for a one-unit change in predictors, e.g., each additional year of age increases charges by about 264.59.

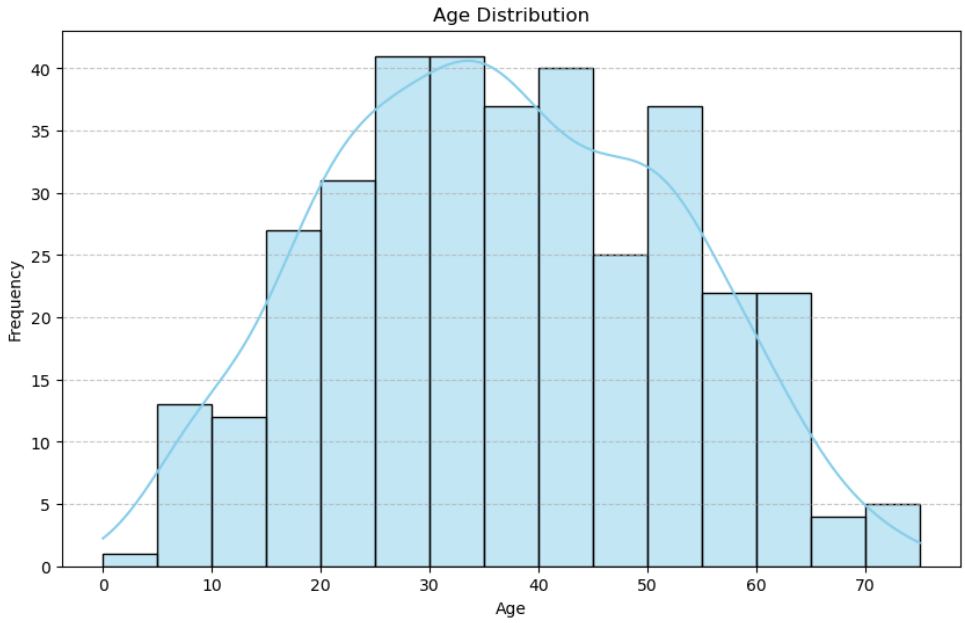
Market segment analysis (Datasets 2)

	erythema	scaling	definite_borders	itching	koebner_phenomenon	polygonal_papules	follicular_papules	oral_mucosal_involvement	knee_and_elbow_involvement	s
0	2	2	0	3	0	0	0	0	1	
1	3	3	3	2	1	0	0	0	1	
2	2	1	2	3	1	3	0	3	0	
3	2	2	2	0	0	0	0	0	3	
4	2	3	2	2	2	2	0	2	0	
...	...	...	...	...	...	...	...	...	...	
361	2	1	1	0	1	0	0	0	0	
362	3	2	1	0	1	0	0	0	0	
363	3	2	2	2	3	2	0	2	0	
364	2	1	3	1	2	3	0	2	0	
365	3	2	2	0	0	0	0	0	3	

366 rows × 35 columns

- The dataset includes a variety of features that describe clinical and possibly histological characteristics, such as **erythema**, **scaling**, **definite borders**, **itching**, **koebner phenomenon**, and more, suggesting a focus on dermatological conditions.
- It also contains a column for **age**, indicating the age of the individuals, which varies widely across the dataset, reflecting a diverse study population.
- The **class** column is used as a categorical target variable, with values from 1 to 6, indicating different conditions or disease states that the dataset aims to classify

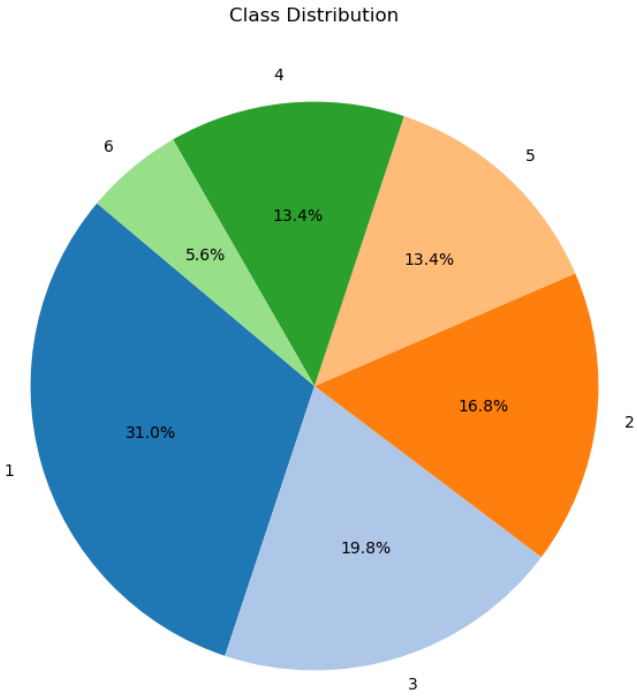
histogram of age distribution

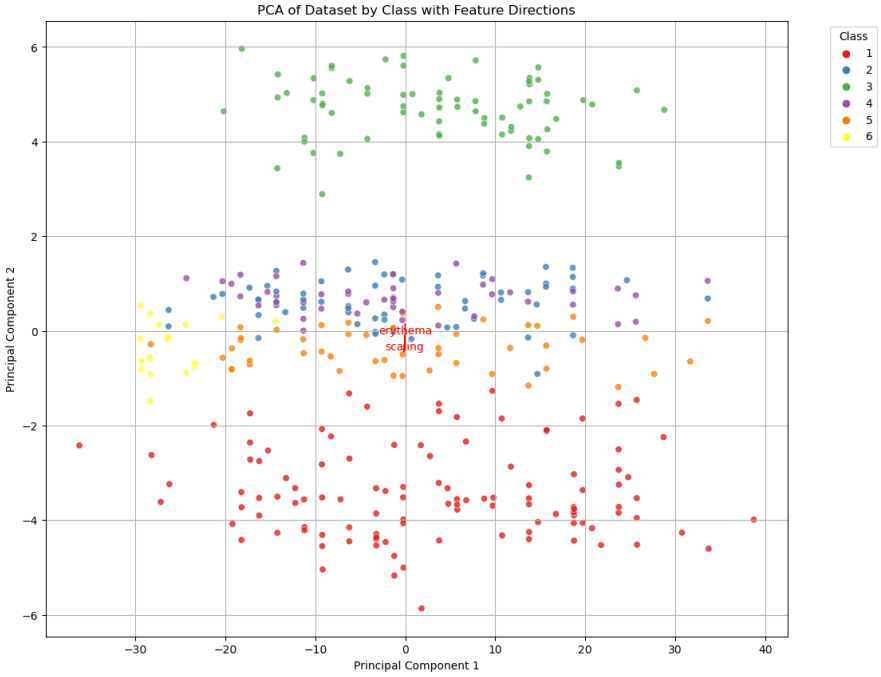


1. **Age Range**: The histogram covers a wide range of ages, from 0 to over 70 years, indicating that the dataset includes a broad demographic.
2. **Peak Age Group**: The highest frequency (mode) seems to be in the age group of around 30 to 40 years, where the tallest bar is located.
3. **Symmetry**: The distribution appears to be fairly symmetric about the peak, with a slight right skew indicated by the longer tail extending toward the older age groups.

The chart shows the distribution of a certain variable across six different classes

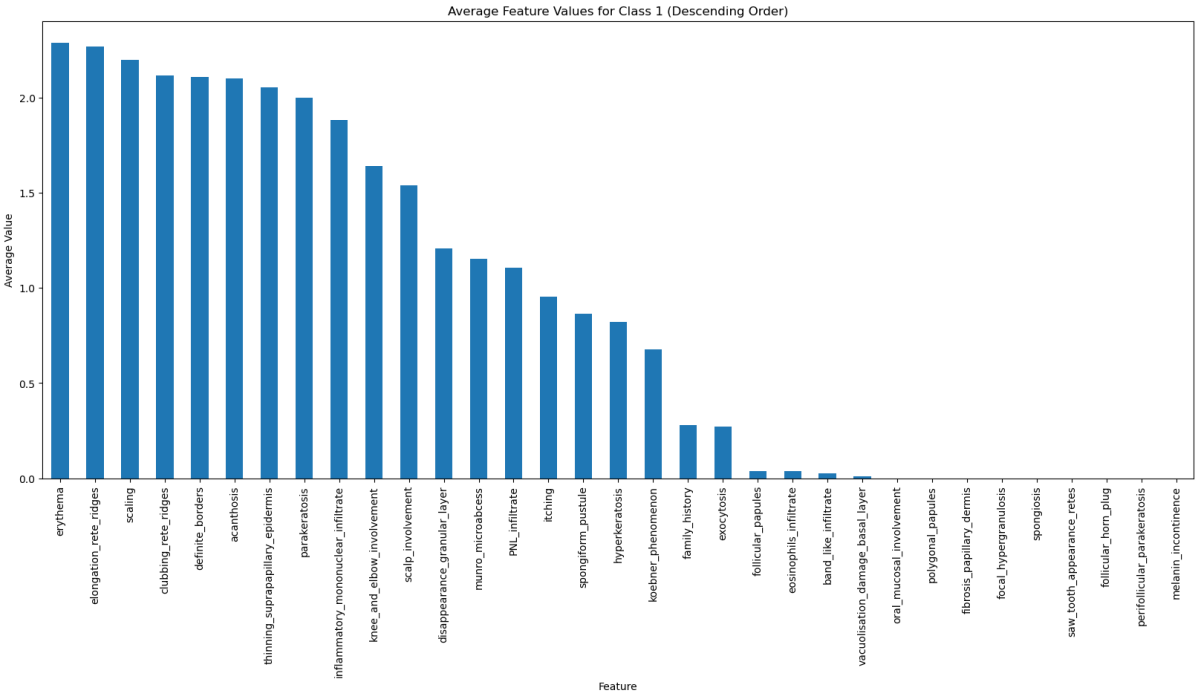
1. **Largest Class**: Class 1 is the largest group, constituting 31.0% of the dataset. This suggests that whatever condition or characteristic Class 1 represents, it is the most common within this data.
2. **Smallest Class**: Class 6 is the smallest group, making up 5.6% of the dataset. This could indicate that the condition or characteristic represented by Class 6 is the least common or least frequently observed in the data.
3. **Balanced Distribution**: Aside from the largest and smallest classes, the remaining classes (2, 3, 4, and 5) are relatively balanced, with each class making up between 13.4% to 19.8% of the data. This suggests a fairly even distribution among these classes.





This plot offers a deeper insight into how the original features influence the separation between classes in the reduced dimensionality space created by PCA.

Bar graph now displays the average values of various features for Class 1



Understanding Class 1: The distribution of feature averages provides insights into the defining characteristics of Class 1. Features with higher averages could be critical in differentiating Class 1 from other classes, potentially corresponding to specific symptoms, signs, or pathological characteristics of the skin condition(s) this class represents.

The age distribution for in Class 1

- **Count**: There are 111 individuals in Class 1.
- **Mean Age**: The average age is approximately 39.38 years..
- **Minimum Age**: The youngest individual in Class 1 is newborn (0 years).
- **25th Percentile**: 25% of individuals are 27 years old or younger.
- **Median Age (50th Percentile):** The median age is 39 years, meaning half of the individuals are younger than 39 and half are older.
- **75th Percentile**: 75% of individuals are 52.5 years old or younger.
- **Maximum Age**: The oldest individual in Class 1 is 75 years old.

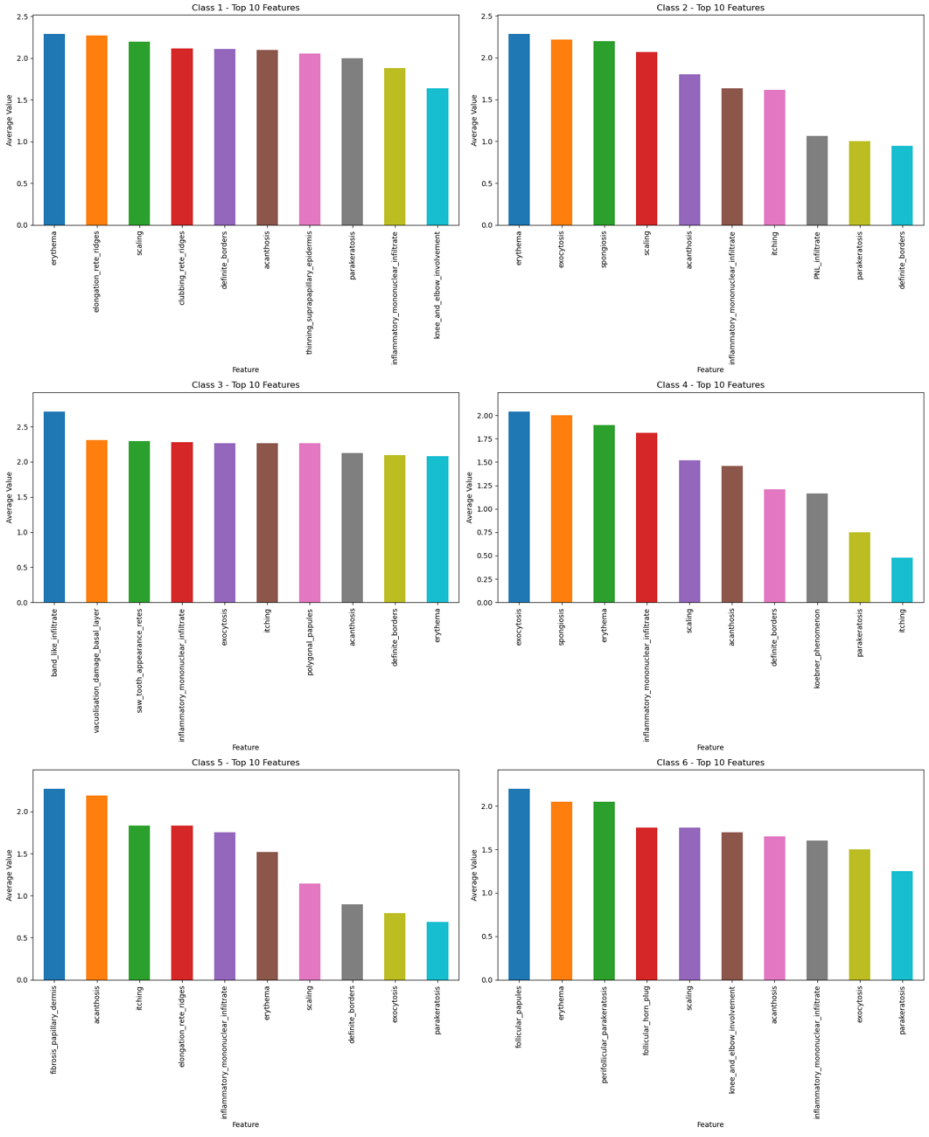
Analysis:

- The age distribution for Class 1 spans a wide range, from newborns to 75 years old, with a fairly even spread across the age spectrum. This suggests that the condition(s) represented by Class 1 can affect individuals at any age.
- The distribution appears to be roughly symmetric around the mean, with a slight skew towards older ages, as indicated by the mean being slightly higher than the median.
- The presence of individuals across all age groups may indicate that the condition(s) associated with Class 1 are not particularly age-specific, affecting a broad demographic.

Top 10 Symptoms from each class

Key Observations Across Classes:

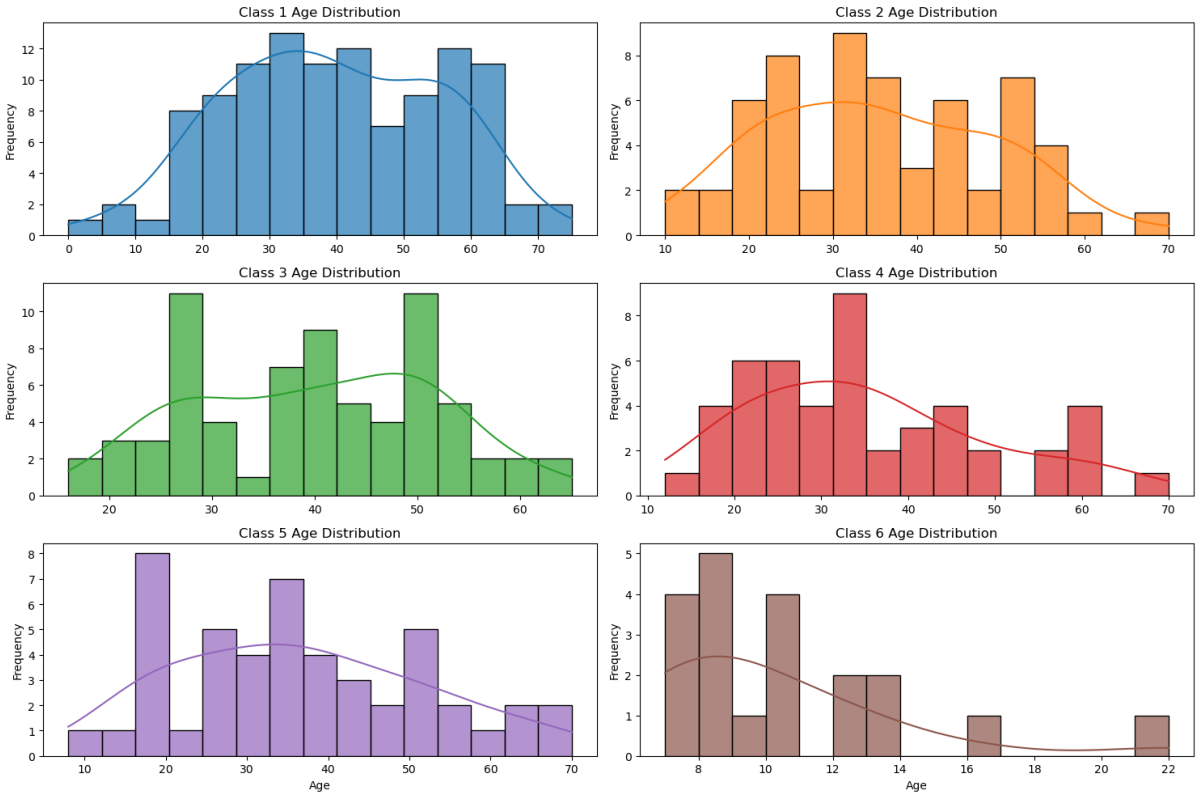
Distinctive symptoms: Each class has a unique set of top symptoms, indicating distinctive characteristics that can help differentiate between the conditions represented by each class.



Common symptoms: While each class has its unique top symptoms, there may be common symptoms that appear across multiple classes, albeit with different average values. This could suggest overlapping symptoms or characteristics among the conditions represented.

Variability in Importance: The importance (average value) of top symptoms varies significantly across classes, underlining the heterogeneity within the dataset. Some features are highly prominent in certain classes but not as significant in others.

The age distribution for all Class 1-6



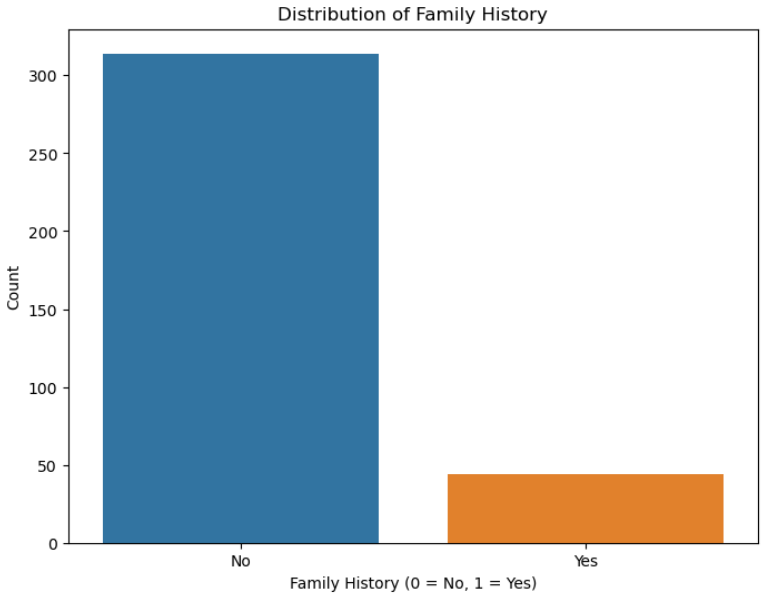
- Classes 1, 2, 4, and 5 cover a broad age range, primarily affecting adults, with a notable concentration in the mid-30s to early 40s. These conditions are prevalent across a wide demographic, suggesting a need for healthcare strategies that cater to a diverse age group.
- Class 3 is more focused on the adult population, particularly middle-aged individuals, indicating conditions with higher prevalence or better diagnosis rates in this demographic.
- Class 6 is unique, targeting children and young adolescents, and highlights the importance of pediatric-specific healthcare interventions.

The distribution of values in the "family_history"

- **87.71%** of individuals have no family history of the condition (value 0).
- **12.29%** of individuals have a family history of the condition (value 1).

This indicates that a majority of the individuals in the dataset do not have a family history of the condition, with only a small fraction reporting a positive family history. This could suggest that for the conditions represented in this dataset, genetic or hereditary factors might play a less dominant role, or at least, are not commonly reported among the cases included in this analysis.

family history of certain health conditions can indeed affect health insurance in several ways



1. **Premium Costs:** Insurance companies may consider an individual's family medical history when determining premium costs. If there's a known genetic predisposition to certain conditions like heart disease, cancer, or diabetes, insurers might view this as a higher risk and could potentially charge higher premiums.
2. **Coverage Terms:** The specific terms of coverage, including exclusions and limitations, might be influenced by family history. For instance, there may be waiting periods for coverage related to conditions for which there is a strong family history.
3. **Pre-existing Condition Clauses:** In some health insurance plans, especially those that are not subject to certain regulations like the Affordable Care Act in the United States, a family history of certain conditions might lead to exclusions for related illnesses as pre-existing conditions.
4. **Risk Assessment:** Insurance underwriters use family medical history as part of their risk assessment process. This can affect eligibility for some types of insurance or result in higher rates.